Interim Analyses in Clinical Trials: Classical vs. ad hoc vs. Bayesian Approaches

Donald A. Berry*

University of Minnesota
School of Statistics

Technical Report No. 418
May 1983

*Research supported in part by NSF grant NSF/MCS 8102477

ABSTRACT

Interim analysis in clinical trials is discussed from the point of view of both classical hypothesis testing and Bayesian inference. The application of P-values in this setting is criticized. The ad hoc approach is recommended over adjusting P-values for the possibility of stopping the trial earlier or later. A Bayesian approach is presented in which stopping occurs when the probability that one treatment is better exceeds a specified value.

Key words and phrases: Clinical trials, interim analysis, P-values, hypothesis testing, Bayesian inference, posterior probability, likelihood principle.

1. Introduction

When a researcher peruses accumulating data from a clinical trial with a possibility of early termination, certain kinds of inferences to be drawn from the trial may be affected. This is well understood by biometricians but not by researchers (McPherson 1982). Readers of medical journals do not ask whether a trial might have been stopped sooner but was not; they take the data at face value. And yet classical Neyman-Pearsonian tests require specifying all possibilities in advance. Important deviations from protocol can make classical inferences impossible. This is one reason, perhaps the most important reason, that sequential designs of clinical trials have been so little used (Simon 1977, Byar et al. 1974).

Biostatisticians have had a substantial impact on this aspect of the design of clinical trials. Dissidents (Anscombe 1963, Cornfield 1966, Weinstein 1974, Berry 1980) have had essentially no effect on statistical practice. This controversial issue is an important one: at stake is human suffering and financial resources of sponsoring agencies and pharmaceutical companies. If some trials last too long and others end prematurely because statisticians disallow looking at accumulating data or continuing beyond the planned trial termination, then the statistical principle on which this is based should have a solid foundation.

A purpose of this paper is to describe implications of classical hypothesis testing for interim analysis. Strict adherence to classical testing is criticized. A less rigid view--one that will offend some

purists but appeal to clinicians--is advocated. The Bayesian approach is discussed.

At the heart of the controversy is the distinction between P-values and probabilities of hypotheses. This issue is discussed in the next two sections; a likelihood approach is discussed in Section 3. The classical approach to interim analysis is discussed in Sections 4 and 5 and compared to a Bayesian approach in Section 5. These are related in Section 6 to an ad hoc approach taken by most researchers. This ad hoc approach uses classical fixed size analyses imbedded in a Bayesian philosophy.

2. P-values and Neyman-Pearson Testing

Most consumers of statistical inference incorrectly regard a P-value as related to the probability of the truth of a null hypothesis. They act as though H_0 is false when the P-value is sufficiently small--usually less than 0.05--feeling that their action is correct with some high probability.

Calculating the probability of H_0 from available data requires the use of Bayes's theorem. This in turn requires a probability assessment of H_0 separate from (or prior to) the data, and pushes many "Bayesians" to adopt a subjective view of probability (Savage 1954). Such blatant subjectivism repels many statisticians. They refuse to pay this price and so reject all Bayesian thought.

P-values are understood by trained statisticians but by few others. For example, two M.D.'s (Diamond and Forrester 1983) asked 24 of their colleagues this multiple choice question: "What would you conclude if a properly conducted, randomized clinical trial of a treatment was reported

to have resulted in a beneficial response ($p < 0.05$)?

1. Having obtained the observed response, the chances are less than 5% that the therapy is not effective.

2. The chances are less than 5% of not having obtained the observed response if the therapy is effective.

3. The chances are less than 5% of having obtained the observed response if the therapy is not effective.

4. None of the above."

The authors say that all responders had difficulty distinguishing the subtle differences between the choices. Of the 24 responders, 11 chose #1 and one other gave an "incorrect" answer. The authors say #3 is correct. However, #3 requires the insertion "or more extreme responses" to be correct. As the choices stand, #4 is correct! Interestingly, 19 of the 24 said they would prefer to know the answer to #1, the (Bayesian) posterior probability; in second place was #2.

Diamond and Forrester (1983) cite a number of important large-scale clinical trials in which a high degree of significance was obtained originally, but the conclusion was later contradicted. Blaming P-values, they make what they call an "arrogant pronouncement:" "The published conclusions of many clinical trials are ill-founded, and may be wrong." They go on to claim that these mistakes would not have been made using a Bayesian approach.

Clinical trials whose results are eventually contradicted do little for the credibility of statistics and statisticians. Regardless of what P-values really mean, consumers will expect that conclusions from clinical

trials (reject H_0, e.g.) are correct with some high probability. This expectation involves posterior probabilities and not P-values. If someone were to find that more than half the conclusions (at $P<0.05$) of all published clinical trials were wrong, this would not be inconsistent with the inferences made from the results of the trials! Disheartening but not inconsistent.

3. Likelihoods vs. Classical Tests

P-values are tail areas. They are calculated by integrating over results more extreme than that obtained, but these more extreme results have not themselves been observed. This characteristic prompted Harold Jeffreys's (1961) criticism: "... a hypothesis which may be true may be rejected because it has not predicted observable results which have not occurred." The size of the tail affects the inference.

As an example, consider the following problem: Ten tosses of a coin result in 2 heads and 8 tails; test the hypothesis that the coin is fair against the one-sided alternative that the probability of heads is less than 1/2. The solution depends critically on the intentions of the tosser, but when confronted with this problem, few people, statisticians included, ask the key question: Why did the tosser choose to stop? If you think this question is irrelevant then you do not take hypothesis testing as seriously as do many statisticians--and you might be a Bayesian at heart!

Solution 1. Suppose the tosser planned ten tosses and would have allowed nothing to stand in the way of this objective. Then the appropriate

distribution is binomial:

$$P = \left[\binom{10}{2} + \binom{10}{1} + \binom{10}{0}\right] \left(\frac{1}{2}\right)^{10} \doteq 0.055.$$

Solution 2. Suppose the tosser planned to obtain two heads and would have tossed as long as necessary. Then the appropriate distribution is negative binomial:

$$P = 9\left(\frac{1}{2}\right)^{10} + 10\left(\frac{1}{2}\right)^{11} + 11\left(\frac{1}{2}\right)^{12} + \ldots \doteq 0.020.$$

Solution 3. Suppose the tosser planned to stop and reject the null hypothesis if there were 0 heads in the first 4 tosses, 1 head after 7 tosses, or 2 after 10--stopping after 10 tosses in any case. Backward induction gives

$$P = 41/512 \doteq 0.080$$

Solution 4. Suppose the tosser planned to continue indefinitely, stopping to reject H_0 as in Solution 3, but also if there were 3 heads in 14 tosses, 4 in 17, etc. An iterative procedure gives

$$P \doteq 0.145.$$

Solution 5. Suppose the tosser planned to stop when dinner was ready and reject H_0 if the proportion of heads was no greater than 0.2. No P-value can be calculated and such data cannot be analyzed legitimately using the Neyman-Pearson approach.

The possibility of different solutions depending on the intentions of the experimenter are counter to the likelihood principle (Cornfield

1966, Barnett 1973). This principle says, roughly speaking, that two sets of data which give rise to the same likelihood function should give rise to the same inferences. In the above problem the likelihood of p, the probability of heads, is proportional to

$$p^2(1-p)^8 \ , \ 0 < p < 1,$$

in all five solutions; the proportionality constant which makes this a probability for the random variable in question varies depending on the solution but is immaterial in the likelihood.

The likelihood principle implies that the question "Why did the tosser stop?" is not relevant for making inferences. (This assumes, of course, that the decision to stop is not a function of p, or of future observations--for example, tossing stopped when the coin was lost and it is more likely to be lost when it is about to land heads.)

The likelihood principle requires a model under which the data are assumed to be produced. If the model is wrong (and it frequently is!) then the inferences can be wrong for that reason. A weaker version which does not rely on a particular likelihood is the following, which is related to the Conditionality Principle (Berger 1980).

Data Principle: Inferences depend only on the data obtained and the experiment performed, and not on data not obtained or on experiments contemplated but not performed.

Objecting to using error probabilities in experimentation, Anscombe (1963) says: "'Sequential analysis' is a hoax. The correct statistical

analysis of the observations consists primarily of quoting the likelihood function. So long as all observations made are fairly reported, the sequential stopping rule that may or may not have been followed is irrelevant. The experimenter should feel entirely uninhibited about continuing or discontinuing his trial, changing his mind about the stopping rule in the middle, etc., because the interpretation of the observations will be based on what was observed, and not on what might have been observed but wasn't."

Since tail areas are not consistent with the likelihood principle, "likelihooders" require another mode of inference. One such is the likelihood ratio test for simple null and simple alternative hypotheses (but not the generalization for compound hypotheses which uses the maximum of the likelihood function in the ratio). Another is the Bayes test, which applies for arbitrary hypotheses. An example is described below.

Data from an actual trial (in which the likelihood $p^2(1-p)^8$ considered above arises at one point) were considered by Cutler, et al. (1966) and Lachin in (Tygstrup, et al. 1982, pp. 241-242). This sequential trial (Freireich 1963) involved two treatments for acute leukemia. Patients were treated in matched pairs and the outcome of interest was time to remission; only the better treatment in each pair was analyzed. The data are shown in Table 1. Using repeated significance testing analysis, $P < 0.05$ was attained with the 18th pair, at which time the remaining three pairs had already been randomized to treatment. (An interesting dilemma presents itself for the classical purist: to what extent should the last three pairs affect the P-value in this approach?--presumably not

at all. But I prefer an approach to inference that allows for different inferences when the last three favor A and when they favor B.)

The results of various analyses are also shown in Table 1. These are described below.

Let $p_B = 1-p_A$ be the probability that treatment B is preferred. Assume that information regarding effectiveness of the two treatments apart from the trial is symmetric. Consider testing H_0: $p_A = r$ vs. H_1: $p_B = r$ (that is, $p_A = 1 - r$) for fixed r. Also require that error probabilities α and β be equal. Wald's sequential likelihood ratio test (SLRT) (Lindgren 1976, pp. 310-317) is to continue as long as

$$c^{-1} < \Lambda < c,$$

where $c = \alpha^{-1} - 1$ and Λ is the likelihood ratio:

$$\Lambda = \frac{r^{n_A}(1-r)^{n_B}}{r^{n_B}(1-r)^{n_A}} = \left(\frac{r}{1-r}\right)^{n_A-n_B};$$

n_A and n_B are the numbers of preferences for A and B respectively. In general, n_A and n_B are jointly sufficient; $n_A - n_B$ is sufficient here because H_0 and H_1 are simple and symmetric. The appropriate P-value is $(1+\Lambda)^{-1}$.

P-values for the SLRT with r = 0.55, 0.6, 0.7, and 0.8 are shown for the accumulating data in Table 1. Significance is reached after 21, 12, 8, and 7 pairs, respectively.

These P-values are also Bayesian probabilities in the following set-up. Suppose the prior probabilities of $p_A = r$ and $p_B = r$ are both 0.5. Then,

Table 1. Analysis of a Sequential Trial Involving Acute Leukemia

Patient Pair	Preferred Treatment	SLRT P-value (r=0.55)	SLRT P-value (r=0.6)	SLRT P-value (r=0.7)	SLRT P-value (r=0.8)	$P(p_B>0.5)$ (Uniform Prior)
1	A	0.45	0.40	0.30	0.20	0.25
2	B	0.50	0.50	0.50	0.50	0.50
3	A	0.45	0.40	0.30	0.20	0.31
4	A	0.40	0.31	0.16	0.059	0.19
5	A	0.35	0.23	0.073	0.015	0.11
6	B	0.40	0.31	0.16	0.059	0.23
7	A	0.35	0.23	0.073	0.015	0.14
8	A	0.31	0.17	0.033	0.0039	0.09
9	A	0.27	0.12	0.014	0.0010	0.055
10	A	0.23	0.081	0.0062	2.4×10^{-4}	0.033
11	A	0.20	0.055	0.0027	6.1×10^{-5}	0.019
12	A	0.17	0.038	0.0011	1.5×10^{-5}	0.011
13	A	0.14	0.025	4.9×10^{-4}	3.8×10^{-6}	0.0065
14	B	0.17	0.038	0.0011	1.5×10^{-5}	0.018
15	A	0.14	0.025	4.9×10^{-4}	3.8×10^{-6}	0.011
16	A	0.12	0.017	2.1×10^{-4}	9.5×10^{-7}	0.0064
17	A	0.099	0.011	9.0×10^{-5}	2.4×10^{-7}	0.0038
18*	A	0.083	0.0076	3.8×10^{-5}	6.0×10^{-8}	0.0022
19	A	0.069	0.0051	1.7×10^{-5}	1.5×10^{-8}	0.0013
20	A	0.057	0.0034	7.1×10^{-6}	3.7×10^{-9}	7.5×10^{-4}
21	A	0.047	0.0023	3.0×10^{-6}	9.3×10^{-10}	4.3×10^{-4}

*Significance ($P<0.05$) reached in actual trial analysis

by Bayes's theorem, the current probability of $p_B = r$ is

$$P(p_B=r|n_A,n_B) = \frac{(0.5)r^{n_B}(1-r)^{n_A}}{(0.5)r^{n_B}(1-r)^{n_A}+(0.5)r^{n_A}(1-r)^{n_B}} = \frac{1}{1+\Lambda},$$

which is the P-value indicated above.

The last column in Table 1 is the probability that treatment B is better than treatment A (that is, $p_B > 0.5$) given the current data when the prior distribution for p_B (and therefore also p_A) is the uniform density on the interval (0,1). From Bayes's theorem,

$$P(p_B>0.5|\text{data}) = \int_0^{\frac{1}{2}} u^{n_A}(1-u)^{n_B}du / \int_0^1 u^{n_A}(1-u)^{n_B}du,$$

an incomplete beta function. (This probability is not a function of n_A-n_B alone.) This becomes less than 0.05 with the tenth pair and approximately 0.01 with the twelfth. The uniform prior may not be unrealistic in view of the following claim by Lachin (Tygstrup 1982, pp. 241-242): "When a group of physicians are shown these results line by line and asked when they would have been ethically compelled to stop the study, usually few would be willing to go beyond the twelfth pair."

4. Interim Analysis; An Example

A trial is planned to compare two treatments; n = 100 paired observations are involved. One interim analysis is planned halfway through the study. If a significant mean difference is detected after 50 pairs then the trial will be stopped and significance proclaimed. Assume the treatment difference is normal with, for convenience, known variance.

Let Z_1 denote the standardized difference after the first 50 pairs and Z_2 the standardized difference for the second 50 pairs. The null hypothesis of no difference is rejected if $|Z_1|>1.96$ or if $|Z_1+Z_2|>1.96\sqrt{2}$ The nominal P-value is 0.05. But the actual rejection probability when the null hypothesis is true is

$$P(|Z_1|>1.96, |Z_1+Z_2|>2.77)$$

which is obviously greater than 0.05 and is given below.

The joint distribution of (Z_1, Z_1+Z_2) under the null hypothesis is normal with mean vector 0 and covariance matrix

$$\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}.$$

So the actual P-value is

$$1 - \int_0^{1.96\sqrt{2}} \int_{-1.96}^{1.96} \frac{1}{\pi} e^{-(u_1^2 - u_1u_2 + u_2^2/2)} du_1 du_2 \doteq .0831,$$

obtained numerically. To have a true level of 0.05 requires a nominal level of 0.0294 (McPherson (1982) obtained 0.0300), corresponding to a critical value of 2.178 for $|Z_1|$ and for $|Z_1+Z_2|/\sqrt{2}$.

Naturally, the nominal level decreases with the number of interim analyses. If, for example, the data are analyzed after each group of ten pairs then the nominal level becomes 0.0105 (McPherson 1982).

The optimal number of interim analyses is discussed in the next section from both classical and Bayesian points of view.

5. The Number of Looks: Classical vs. Bayesian

The benefit of a sequential or "group sequential" approach is the possibility that a small amount of data will be conclusive and allow for early termination of the trial. This is an important consideration when observations are dear. Ethical considerations when human lives are involved have been discussed extensively (Chalmers 1982, Weinstein 1974, Winfrey 1978). As another kind of example, consider a Phase I drug comparison trial in which myocardial infarctions are artificially induced in dogs, after which extensive experimental procedures are followed and measurements made. Such a trial can easily cost more than $1000 per dog.

Minimizing expected costs in such a trial (subject to obtaining conclusive evidence) is equivalent to minimizing average sample size (ASN). But adjusting P-values can make the ASN increase as the number of looks increase, with the minimum ASN occurring in the fixed sample size case.

For example, McPherson (1982) considers normal sampling (as in Section 4) for testing the unknown mean $\delta = 0$. The ASN depends on δ as well as the number of looks. To obtain a single measure of sample size McPherson averages the ASN with respect to a prior distribution on δ (thereby combining an otherwise classical approach with a Bayesian notion). For four different prior distributions, maximal reductions in ASN over fixed sample sizes vary from 0 to about 30% with greatest savings occurring when both large and small deviations from $\delta = 0$ are likely a priori. In these examples the optimal number of looks varies from 1 to 8, with ASN increasing for numbers of looks greater than the optimum.

In a full Bayesian approach, the ASN decreases indefinitely with the number of looks. The minimum occurs with a fully sequential approach in which accumulating data are constantly monitored.

In analogy with McPherson (1982), assume the observations X_i are $N(\delta, 1)$. Let the prior distribution for δ be $N(0, \tau^2)$; this family includes two of McPherson's priors mentioned above, namely $\tau = 0.1$ and 0.2. After observing $X_1, \ldots, X_n$, letting $S_n = \Sigma_1^n X_i = n\bar{X}$, the posterior distribution of δ is

$$N\left(\frac{n\bar{X}}{n+\tau^{-2}}, \frac{1}{n+\tau^{-2}}\right)$$

(DeGroot 1970).

Assume the data are to be analyzed in groups of size g; g = 1 is the fully sequential case. Initially, $P(\delta>0) = P(\delta<0) = 0.5$. The trial will be stopped after n observations if either treatment has been shown to be effective with sufficiently high probability: when either $P(\delta>0|X_1, \ldots, X_n)$ or $P(\delta<0|X_1, \ldots, X_n)$ is greater than, say, 0.90 or 0.95. Since sampling can go on indefinitely in this setting--in fact, the expected sample size is infinite--for any g and finite τ, truncation is necessary. Sampling will be terminated (if it has not stopped previously) whenever the current distribution of δ has standard deviation sufficiently small, say less than 0.05. Since, in addition, $P(\delta>0|X_1, \ldots, X_n)$ will be moderate in size when the truncation point is reached, the distribution of δ will be heavily concentrated near 0, indicating that neither treatment is very much better than the other. The conclusion concerning the sign of δ will not be as strong as when sampling is terminated before the truncation point. But it is legitimate to calculate $P(\delta>0|X_1, \ldots, X_n)$ or any other characteristic of

the posterior distribution regardless of the reason for stopping. (Actually, since the posterior depends on the prior, it is more appropriate to give the posterior for various priors, allowing the consumer to specify the prior.)

The requirement that sampling stops when $(n+\tau^{-2})^{-\frac{1}{2}} \leq 0.05$ is equivalent to $n \geq 400 - \tau^{-2}$. The quantity $n + \tau^{-2}$ is the sum of the actual and "prior" sample sizes; sampling is terminated whenever this "effective sample size" is at least 400.

There are at least two ways of simulating to find the distribution of the sample size, or its mean, using the stopping rule described above. One is to find the distribution as a function of δ by simulating over a grid of values; then average with respect to the prior distribution of δ. Another is to increment S_n with X_{n+1} at the n^{th} stage using the conditional distribution of S_{n+1} given S_n:

$$N\left(\frac{n+1+\tau^{-2}}{n+\tau^{-2}} S_n, \frac{n+1+\tau^{-2}}{n+\tau^{-2}}\right)$$

Table 2 shows the expected sample sizes evaluated using the latter approach, for various τ and g. Evidently, great savings in sample size are possible.

The average sample size is greatest in the fixed sample size case and least when the data are continually monitored. However, there are few trials in which the response to treatment is immediate. For this and other logistical reasons, continual monitoring is impossible or impractical. The message here is that the data should be analyzed as frequently as possible; nothing is lost in inferential ability when the data are analyzed often,

Table 2. Expected Sample Size with Maximum of $400\text{-}\tau^{-2}$ Paired Observations (and Standard Deviation)

Number of groups† (Approx.)	Group size g† (Approx.)	τ=0.2 90%	τ=0.2 95%	τ=0.5 90%	τ=0.5 95%	τ=1.0 90%	τ=1.0 95%	τ=2.0 90%	τ=2.0 95%
1	400	375 (0)	375 (0)	396 (0)	396 (0)	399 (0)	399 (0)	400 (0)	400 (0)
2	200	316 (87)	337 (76)	328 (96)	352 (89)	327 (96)	351 (86)	327 (96)	351 (86)
3	133	280 (110)	309 (99)	282 (120)	316 (111)	281 (121)	315 (112)	281 (121)	315 (112)
4	100	261 (121)	295 (112)	253 (131)	293 (126)	250 (132)	292 (126)	249 (132)	291 (127)
5	80	235 (125)	275 (118)	228 (136)	269 (134)	224 (135)	267 (135)	223 (135)	265 (135)
10	40	190 (134)	238 (146)	160 (135)	209 (146)	153 (134)	203 (147)	151 (132)	203 (148)
15	27	167 (131)	215 (135)	130 (130)	177 (148)	122 (127)	168 (146)	117 (125)	167 (147)
20	20	156 (126)	199 (130)	111 (122)	158 (144)	102 (119)	148 (143)	98 (116)	145 (142)
25	16	144 (127)	194 (136)	101 (119)	143 (141)	89 (112)	132 (139)	85 (109)	128 (138)
50	8	127 (122)	175 (134)	70 (103)	108 (130)	58 (94)	91 (123)	53 (88)	86 (121)
100	4	117 (121)	165 (136)	52 (88)	88 (120)	37 (76)	65 (107)	33 (70)	60 (104)
200	2	103* (114)	152* (133)	44* (82)	68* (104)	24 (59)	47 (93)	20 (54)	40 (89)
400	1	104** (116)	155** (135)	38** (77)	65** (106)	18 (48)	37 (81)	13 (41)	26 (69)

Based on 5000 simulations (s.e. approx. 0.014 s.d.) except
* 1500 simulations (s.e. approx. 0.026 s.d.)
** 1000 simulations (s.e. approx. 0.032 s.d.)
† These may be smaller, depending on τ. For example, since 25 x 15 = 375, 25 groups of size 15 are sufficient where τ = 0.2.

and smaller sample sizes will frequently result. (This does not mean, of course, that the clinician can do this frequent monitoring; the rationale for double-blinded studies is strong and does not depend on one's approach to statistical inference.)

There are two main objections by classical statisticians to a Bayesian analysis. One is the difficulty and arbitrariness in picking a prior distribution. The other is the possibility of "sampling to a foregone conclusion." The first is a legitimate complaint; the second is not. Sampling to a foregone conclusion is possible in the classical set-up when one considers Type I and Type II errors. But posterior probabilities do not behave like error probabilities. For example, suppose $X_1, \ldots, X_n$ have been observed. The probability of $\delta>0$ given $X_1, \ldots, X_n, X_{n+1}$ is a random variable when conditioning on $X_1, \ldots, X_n$. Its expected value is precisely the current probability of $\delta>0$; that is, the probability of $\delta>0$ is a martingale (unlike P-values). So if the current probability is 0.94, it can increase to above 0.95 with the next observation or it may decrease. In the case of normal sampling described above, the expected number of observations required to convert a current probability of 0.94 into one greater than 0.95 is infinite.

6. Classical vs. Bayesian vs. ad hoc

Medical researchers may not be Bayesians in any formal sense, but they act like Bayesians in many ways; in particular, in their lack of inhibition regarding interim analysis. Few reports of clinical trials that I have seen adjust P-values, even though data monitoring with the possibility of early termination has taken place. According to McPherson (1982): "... such

arguments [for adjusting P-values] appear not to have captured the imagination of the great majority of investigators." In fact, reports of ongoing clinical trials are published with P-values calculated as though the trial was fixed in size to be the current size.

Consider the following scenario. A medical researcher adopts a sequential scheme--with normal observations, say. The researcher decides to entertain up to 10 stages, stopping before with a two-sided significance level of 0.05. The nominal P-value required is 0.0105. At the first stage, $Z = 2.559$ ($P = 0.0105$) and so the trial was stopped, with the researcher proclaiming $P = 0.05$.

While the results are being presented at a professional meeting, a representative from the sponsoring agency states that the researcher was not to receive funding beyond the first stage in any case. Does this mean that, since the original intentions could not have been carried out, P is actually 0.0105?

Someone from another agency is in the audience and gets up to say that they knew about the trial and the first agency's plan to withdraw sponsorship. It seems they were willing to take over sponsorship for one more stage should that have been necessary. Now, P is neither 0.0105 nor 0.05, but 0.0185.

On the other hand, if the second agency had said they would have taken over sponsorship only if the investigator would have agreed to up to 20 stages, then the P-value becomes >0.05 and the results are not significant! Unless of course the researcher says he might not have agreed. What then is the P-value?

The point of this story is to suggest that it is difficult if not impossible to adher strictly to a classical approach. (Even more, as Berger (1980, p. 354) says, "In a very strict sense, one wonders how the classical statistician can do any analysis whatsoever.")

P-values (calculated assuming fixed sample sizes) may be reasonable as measures of extremity, answering the question "How unusual are the data?", but they should not be taken too seriously. The casual approach employed by most researchers of reporting such P-values for the various measurements (efficacy, safety, and side effects) is not without its dangers. Consumers have to be educated to not necessarily act as though H_0 is false simply because P is small. But adjusting P-values is even less desirable, at least in part because it is so poorly understood by the consumer. If the results of a trial are published then the reader should be able to duplicate calculation of the given P-value without having to know what the investigator would have done should various contingencies have arisen.

Indeed, no investigator can specify all possible contingencies. The statistician and the mode of inference must be flexible enough to handle unforeseen developments. A remark of Cornfield in (Cutler, et al. 1966) made in another context is appropriate here:

> Of course a re-examination in the light of results of the assumptions on which the pre-observational partition of the sample space was based would be regarded in some circles as bad statistics. It would, however, be widely regarded as good science. I do not believe that anything that is good science can be bad statistics, and conclude my remarks with the hope that there are no statisticians so inflexible as

to decline to analyze an honest body of scientific data simply because it fails to conform to some favored theoretical scheme. If there are such, however, clinical trials, in my opinion, are not for them.

Acknowledgement. Chih-Kung Wu helped with the simulations reported in Table 2.

REFERENCES

Anscombe, F. J., 1963. Sequential medical trials, J. Amer. Statist. Assoc. 58, 365-383.

Barnett, V., 1973. Comparative Statistical Inference. John Wiley and Sons, London.

Berger, J. O., 1980. Statistical Decision Theory, Foundations, Concepts, and Methods. Springer-Verlag, New York.

Berry, D. A., 1980. Statistical inference and the design of clinical trials. Biomedicine 32, 4-7.

Byar, D. P., Simon, R. M., Friedewald, W. T., Schlesselman, J. J., DeMets, D. L., Ellenberg, J. H., Gail, M. H., and Ware, J. H., 1976. Randomized clinical trials. New Engl. J. Med. 294, 74-80.

Cornfield, J., 1966. Sequential trials, sequential analysis and the likelihood principle. The Amer. Statist. 20, 18-23

Cutler, S. J., Greenhouse, S. W., Cornfield, J. and Schneiderman, M. A., 1966. The role of hypothesis testing in clinical trials. J. Chron. Dis. 19, 857-882.

DeGroot, M. H., 1970. Optimal Statistical Decisions. McGraw-Hill, Inc., New York.

Diamond, G. A. and Forrester, J. S., 1983. Clinical trials and statistical verdicts: Probable cause for appeal. Ann. Internal Med. 98, 385-394.

Freireich, E. J., Gehan, E., Frei, E., Schroeder, L. R., Wolman, I. J., Anbari, R., Burgert, E. O., Mills, S. D., Pinkel, D., Selawry, O. S., Moon, J. H., Gendel, B. R., Spurr, C. L., Storrs, R., Haurani, F., Hoogstraten, B., and Lee, S., 1963. The effect of 6-Mercaptopurine on the duration of steroid-induced remissions in acute leukemia: A model for evaluation of other potentially useful therapy. Blood 21, 699-716.

Jeffreys, H., 1961. Theory of Probability (3rd Ed.). Oxford University Press, London.

Lindgren, B. W., 1976. Statistical Theory, 3rd Ed. Macmillan Publishing Co., Inc., New York.

McPherson, K., 1982. On choosing the number of interim analyses in clinical trials. Statistics in Medicine 1, 25-36.

Savage, L. J., 1954. The Foundations of Statistics. John Wiley and Sons, Inc., New York. (1972 edition, Dover Publications, New York.)

Simon, R., 1977. Adaptive treatment assignment methods and clinical trials. Biometrics 33, 743-749.

Tygstrup, N., Lachin, J. M., Juhl, E. (eds.), 1982. The Randomized Clinical Trial and Therapeutic Decisions, Marcel Dekker, Inc., New York.

Weinstein, M. C., 1974. Allocations of subjects in medical experiments. New Engl. J. Med. 291, 1278-1285.

Winfrey, C., 1978 (March 5). The New York Times.